UDACITY

‹ Back to Data Scientist Nanodegree

Identify Customer Segments

REVIEW

HISTORY

Meets Specifications

Great Work!
Congratulations on completing your project!

- I certainly enjoyed walking through your code. It's very clean and very well commented. I can clearly see the effort that has been put into this.
- You have perfectly taken care of the feedback and incorporated them flawlessly in this attempt.

Here are some additional links to further your knowledge:
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
https://www.kaggle.com/dansbecker/handling-missing-values/notebook
https://datascienceplus.com/find-your-best-customers-with-customer-segmentation-in-python/
https://medium.com/@staceyferreira/a-look-at-customer-segmentation-43e053a8cef1.

Good Luck in your Data Science journey😄

Preprocessing

All missing values have been re-encoded in a consistent way as NaNs.

- Also, add docstring to this function.

Columns with a large amount of missing values have been removed from the analysis. Patterns in missing values have been identified between other columns.

The data has been split into two parts based on how much data is missing from each row. The subsets have been compared to see if they are qualitatively different from one another.

- Good job implementing the previous suggestion.

Categorical features have been explored and handled based on if they are binary or multi-level.

`OST_WEST_KZ` was encoded correctly.

Mixed-type features have been explored, resulting in re-engineered features.

Dataset includes all original features with appropriate data types and re-engineered features. Features that are not formatted for further analysis have been excluded.

A function applying pre-processing operations has been created, so that the same steps can be applied to the general and customer demographics alike.

- You could also add a cell to test your clean function on a small subset of azdias.

Feature Transformation

Feature scaling has been properly applied to the demographics data. Imputation has been performed to remove remaining missing values.

Principal component analysis has been applied to the data to create transformed features. A variability analysis has been performed to justify a decision on the number of features to retain.

Weights on at least three principal components are used to make inferences on correlations between original features of the data. General meanings are ascribed to principal components where applicable.

Clustering

Multiple cluster counts have been tested on the general demographics data, and the average point-centroid distances have been reported. A decision on the number of clusters to use is made and justified.

- Analysis has been corrected. Next you could try using MiniBatchKmeans also for better performance.

Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data.

A comparison is made between the general population and customers to identify segments of the population that are central to the sales company's base as well as those that are not.

- Awesome job including the many Nan rows and treating it like a separate cluster.

DOWNLOAD PROJECT

RETURN TO PATH